THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July-August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Toward Bell System Applications of Automatic Speech Recognition

By J. E. HOLMGREN*

(Manuscript received July 30, 1982)

Advances over the past few years in the field of Automatic Speech Recognition (ASR) have brought more attention to potential Bell System applications of this technology. Before reaching the point of ASR implementation, several human factors problems have to be overcome. This paper describes the central human factors issues, then summarizes the initial steps at Bell Laboratories in attempting to deal with those issues. Findings from observations of customers speaking credit card numbers to operators are described, followed by summaries of three studies investigating control of the speech of ASR system users.

I. INTRODUCTION

The potential of Automatic Speech Recognition (ASR) in telecommunication applications has long been recognized,[1] but until recently the state of ASR technology did not warrant effort beyond the existing laboratory activity. In the past few years, ASR has emerged from the laboratory into commercial use. Several systems are now available providing speaker-dependent word recognition.[2] Such a system must be trained by each user before it can recognize that user's speech.

* American Bell.

Most of these systems require isolated speech in the sense that each vocabulary item must be spoken with a short silent interval preceding and following each word, althougb a few systems allowing connected speech are now available.[2] Speaker-dependent systems have been applied primarily in industrial settings that require hands-free data entry, such as inspection and quality control.

Speaker-dependent ASR systems are inappropriate for many attractive telecommunications applications. Instead, speaker-independent systems, which need no prior training by users, are required. Various network operator services, such as credit card calling and directory assistance, are examples of applications in this category. Through the use of ASR, such services could be automated and used from any telephone that has access to the existing network. A few isolated-speech, speaker-independent ASR systems are already in use in tbe United States and Japan. They provide service to selected groups of users in applications sucb as private network call-routing and banking. Laboratory work is under way to develop a connected-speech, speaker-independent ASR capability.[3]

The use of ASR in universally accessible services raises many human factors questions. This paper summarizes Bell Laboratories initial buman factors work leading toward tbe first network applications of speaker-independent ASR. Our work began in the context of considering the use of ASR in one particular network service: credit card calling. Today, most credit card calls require giving a credit card number (CCN) to an operator. A new service known as Calling Card Service (CCS) automates the handling of credit card calls from *Touch-Tone** telephones by allowing customers to enter their CCNs on tbe *Touch-Tone* telephone number pad. However, calls from rotary telephones must still be handled by operators. ASR would allow automation of all credit card calls.

Our first step in investigating the credit card application was to identify the critical human factors issues that require attention. Several of these issues are common to almost all potential network applications of speaker-independent ASR. A description of the common issues provides perspective on our subsequent human factors work.

II. THE CENTRAL HUMAN FACTORS ISSUES

Issues that surround the user-system dialog were selected as the focus of the work reported here. Although the other issues are only touched on in this paper, they are no less important to any successful network application of ASR.

* Trademark of AT&T.

2.1 User-system dialog

For the foreseeable future, ASR systems will be limited by several aspects of human speech, such as the vocabulary they can recognize, the maximum rate of speech they can handle, their ability to ignore extraneous words and sounds, and the accuracy of recognition per spoken vocabulary item. Thus, care must be taken in the design of any user-system dialog to overcome these limitations.

2.1.1 Instructions

Appropriate instructions are needed to control the speaking rate and vocabulary that untrained speakers use when they encounter an ASR system for the first time. These instructions also must allow experienced users to proceed without unnecessary delay.

2.1.2 Feedback

While current ASR systems are achieving impressive recognition accuracy for limited vocabularies, their accuracy is less than that of a human listener. Therefore, many applications, particularly those where errors are costly, require feedback to the user to ensure correct recognition by the ASR system. For many attractive telecommunications applications of ASR, user input will consist primarily of strings of digits (e.g., telephone numbers or credit card numbers). Several digit feedback provisions are possible. For instance, feedback could be given after each digit, after each group of digits, or after entry of the entire number. The optimal method depends on the nature of the application and the recognition accuracy of the ASR system.

2.1.3 Error correction

Methods are needed that allow users to correct both their own speaking errors and, when feedback is given, recognition errors on the part of the ASR system. Again, several options are available and the optimal method for each application is unclear.

2.1.4 Problem speakers

No matter how good an ASR system may be, there will always be some speakers whose speech cannot be reliably recognized by the system. Any service incorporating ASR will have to provide for some type of alternate treatment for such individuals; this will often mean transfer to an operator or attendant. Detection of problem speakers early in the dialog and swift alternate treatment will be necessary to ensure both service efficiency and customer satisfaction. The best way to detect these speakers has not yet been established.

2.2 Isolated vs. connected speech

Connected speech is preferable to isolated speech for use as input to an ASR system. However, when speaker-independent ASR systems that accept connected speech become available for use, isolated input will still be more accurately recognized. For this reason, in any application of ASR it will be necessary to decide whether the trade-off between the greater ease of use of connected speech and the greater accuracy of isolated speech favors the former or the latter.

2.3 Vocabulary choice and expansion

User and system considerations may often conflict when a vocabulary is selected for any given application of ASR. The most appropriate vocabulary for the speaker may be particularly difficult for the ASR system. For instance, while spoken, spelled input might be a natural way to specify a name to a directory assistance system, the spoken alphabet is a singularly difficult vocabulary for any ASR system to accept.[4] Words in a vocabulary such as the international word-spelling alphabet (Alpha, Bravo, Charlie, etc.) would be much more accurately recognized by machine, but much less convenient for most users. Closely related to vocabulary selection is the problem of vocabulary expansion. Careful selection of new vocabulary items is necessary because adding new words to a vocabulary may change the system's performance on the original set of words.

2.4 Integration of Touch-Tone service with ASR

For any network service using ASR, a large percentage of the customers will be calling from *Touch-Tone* telephones. Thus, it is necessary to consider the possibility of mixed *Touch-Tone* telephone and voice input to an automated service. This raises several questions regarding integration of the two, such as whether to provide both input options at every point in a service, whether to encourage the use of one option over the other, how to make voice input compatible in some sense with *Touch-Tone* telephone input when both are available, etc.

2.5 Template construction

In an ASR system templates represent the words to be recognized in a given application. Template construction is in many respects the most critical hurdle in applying speaker-independent ASR because it is largely the quality of those templates that determines the recognition accuracy of the system across the population of users.[5] Template considerations are included here because of the human factors problems involved in building the speech database needed to construct them.

2.6 System evaluation

The performance of a speaker-independent ASR system in any application depends not only on the characteristics of the system but also on the vocabulary used in the application, the set of templates constructed for that vocabulary, the transmission conditions, and the characteristics of the spoken input. Therefore, evaluating the adequacy of a system in an application involves more than simply obtaining some overall measure of recognition accuracy. Information will be needed about variation in recognition accuracy across segments of the user population, across vocabularies, and across transmission conditions. Other critical aspects of performance will be system response time and the rate of false recognition for words outside the application vocabulary.

III. TSPS OBSERVATION STUDY

To study user-system dialog issues in the application of ASR to credit card calling we gathered data from customers speaking their credit card numbers to Traffic Service Position System (TSPS) operators, who handle all nonautomated credit card traffic. Such data were needed to identify any customer behavior changes necessary to interact successfully with an ASR system.

The particular aspects of customer speaking behavior that we investigated were:

1. Customer segmenting of CCNs. Segmenting, as used here, refers to the tendency of most customers to break a CCN into spoken segments by pausing briefly after speaking a group of three or four digits. Segmenting is important because the per-item recognition accuracy of speaker-independent ASR systems that can accept connected speech is likely to be highly sensitive to the number of items in a connected sequence.
2. Customer vocabulary (e.g., "zero" versus "oh," "hundred" versus "zero zero," etc.).
3. Occurrence of words or sounds other than those used to give the CCN.
4. Frequency of customer mistakes in speaking the CCN and spontaneous correction of those mistakes.
5. Frequency of operator requests for repetition of a portion or all of the CCN. Our interest in this stems from the fact that reliable system performance is possible only when the customer corrects system-recognition errors using feedback from the ASR system; thus, it is useful to have information on the current frequency of operator-requested repetitions.
6. Speaking rates distribution for current customers.

3.1 Summary of results

3.1.1 Number of observations

A total of 3040 credit card calls were observed at three different TSPS offices, 1157 in Milwaukee, 742 in Louisville, and 1141 in Boston. The relatively low number of observations for Louisville reflects a low overall credit card call volume during the observation period. Since the card numbers for the observed calls at each site were not recorded, the number of distinct CCNs among the 3040 calls is not known.

3.1.2 Segmentation of the spoken CCN

The majority of CCNs were spoken in consecutive segments of 3, 3, 4, and 4 digits. Of all observations, 69 percent fell in this category. This is not surprising, since the format of most CCNs is NPA NXX XXXX XXXX and in the three operating companies visited the number is printed on the credit card with that segmenting.

The next most common segmentation was 3 3 4 3 1, used in 17 percent of the observations. The frequency of this segmentation probably reflects carryover from the previous 10-digit-plus-one-letter format used for CCNs up until two months before data collection. In most cases, the new 14-digit numbers were constructed from existing CCNs by adding an initial NPA and changing the terminal letter to a digit. Tbus, despite the fact that the new number was printed on the credit card with the 3 3 4 4 segmentation, customers accustomed to speaking their old number with a 3 4 3 1 segmentation carried that babit over to the new number.

Only four percent of the observed numbers were spoken witb the digits run together. A number was classified as run together if five or more digits were spoken without an intervening pause. This was the third most common segmentation category. The remaining 10 percent of the calls were distributed among several infrequently occurring segmentation categories.

3.1.3 Variation in vocabulary

Regarding the use of "zero" or "oh" wben speaking the digit zero, 29 percent of the 3040 spoken CCNs contained a spoken "zero," 51 percent contained a spoken "oh," 9 percent contained both "zero" and "oh," and 11 percent contained neither. It is interesting that customers frequently say both "zero" and "oh" while speaking a single CCN. While figures are not available on the proportion of observed CCNs that contain multiple occurrences of the digit zero, it is evident tbat the percentage of multiple zero calls involving the use of both "zero" and "oh" is considerably greater than nine percent.

When speaking the CCN, customers used words outside the single-

digit vocabulary in only eight percent of all calls. Almost all of these calls involved some combination of three types of multiple-digit utterances. The most common type was a two-digit combination such as "fifty-six," "eighty-eight," "thirteen," etc. The next most common type was the phrase "double zero" or "double oh" for the digit combination 00, usually when the 00 was the leading pair of a segment. The word "hundred," usually used for a terminal 00 in a segment, was the third most common type.

3.1.4 Extraneous vocalizations

There were relatively few occurrences of extraneous vocalizations from the customer. Less than two percent of all calls included such events. The most commonly used word was "dash," spoken between segments of the CCN. This occurred because some operating companies print the CCN on the credit card with dashes between segments.

3.1.5 Customer correction of errors

The customer corrected an error in the spoken CCN before any request for repetition from the operator on less than three percent of all calls. The most common error was to leave off the area code when giving the CCN. Since the requirement to give the area code had been in force for less than three months at the time these data were collected, customers were still adjusting to the new CCN format.

3.1.6 Operator requests for repetition

The operator asked the customer to repeat some or all of the CCN on less than six percent of all calls. Usually the operator requested repetition of all 14 digits. This occurred because the operator can collect any number of digits entered on the TSPS console number pad only by cancelling all entered digits and beginning again. The three most common reasons for requesting repetition were operator keying errors, the operator misunderstanding or not hearing the customer, and the customer forgetting to give the area code.

3.1.7 Customer speaking rates

Customer speaking rates are of interest only for those calls on which the CCN was entered in a manner otherwise consistent with the constraints likely to be placed on users of any connected-speech, speaker-independent ASR system. Examining these calls allows one to determine if credit card customers were speaking too quickly for ASR systems when their spoken input was acceptable in all other respects. A call was classified as acceptable if it had the following characteristics:

1. Only single-digit words were spoken.

2. No extraneous vocalizations occurred.
3. The customer did not make an error in entry.
4. The operator did not request a repetition.

An operator may request repetition of a CCN, even though the input was acceptable to an ASR system. However, to be conservative, these calls were excluded from the acceptable set, which contained 80 percent of all calls.

As a function of location, the mean speaking time for a CCN is 5.16 seconds in Milwaukee, 5.33 seconds in Louisville, and 5.09 seconds in Boston. Although the variation among these means is small, an analysis of variance shows it to be statistically significant ($F = 11.79$, $p < 0.001$). The mean speaking times for males and females are nearly identical, being 5.17 seconds and 5.19 seconds, respectively.

The performance of connected-speech, speaker-independent ASR systems is likely to fall off rapidly as the rate of speaking connected strings increases beyond about 2.5 words per second.[6] For this reason, speaking rates within connected strings of digits are of more interest than total speaking times for CCNs. To compute a speaking rate for each call, the total speaking time was corrected for the pauses occurring between segments. This correction was made by assuming that for any speaker, the duration of such pauses was roughly equal to that speaker's mean speaking time per digit. Based on the listening experience of the two data collectors, this appears to be a reasonable assumption. The appropriate estimate of speaking rate for a call is therefore given by the following expression:

$$\text{Rate} = (14 + s)/t,$$

where s is the number of intersegment pauses and t is the total speaking time for the CCN. The value of s was based on the segmentation judgment made by each data collector on each call. Calls for which the digits were judged to be run together present problems for this rate estimate. Any call having five or more digits run together should fall in this category, so the number of intersegmental pauses is not constant within the category. For simplicity, it was assumed (based on the data collectors' observations) that the average number of pauses in a call classified as run together was one, so s was set to that value.

The distribution of speaking rates is given in Fig. 1, where, for 94.2 percent of the acceptable calls, the speaking rate was greater than 2.5 digits per second.

3.2 Implications for customer behavior

The above findings indicate two areas where the modal customer-speaking behavior is likely to require modification before customers

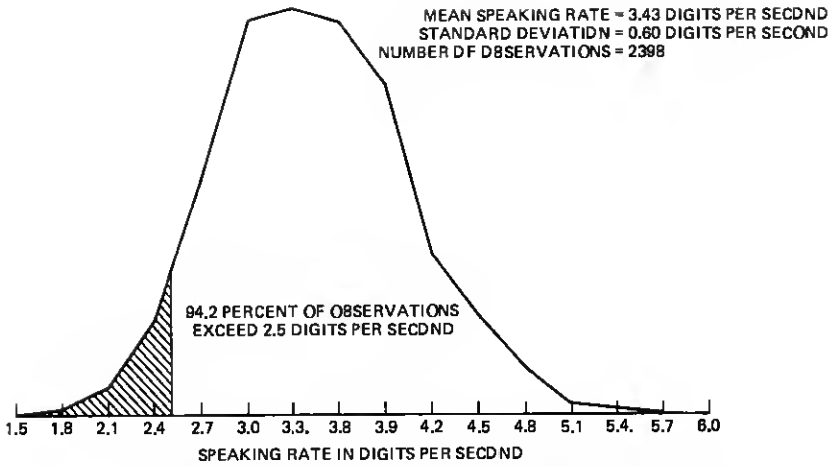

Fig. 1—Distribution of speaking rate for calls classified as acceptable.

could successfully interact with a speaker-independent, connected-speech ASR system. First, assuming that any such system will have considerably more difficulty recognizing "oh" than "zero" (due to the relatively small amount of energy in a spoken "oh" and to coarticulation effects), customers may have to learn to say "zero" in place of "oh". Second, assuming that 2.5 digits per second represent an approximate maximum acceptable speaking rate, most customers will have to lower their rates. Some means is needed of getting all customers under the critical rate without lowering the rate for many of them too much. In other words, both the mean and variance of the speaking rate distribution need to be decreased.

While these data provide some useful baseline information about spoken digit strings in a field environment, generalization of the findings to other situations must be done carefully. The CCN is a long, highly familiar number and credit card customers are experienced callers. Input characteristics may differ for sborter, less familiar digit strings or for a less experienced user population. Also, there are several important questions that could not be addressed in this study. For instance, while there were no important differences among locations on any of the variables investigated, the possibility remains tbat variation in speaker accent will present critical problems for any ASR system. Questions such as this cannot be answered without testing a system on actual speech samples from a field environment.

Despite its limited scope, this study was useful in guiding subsequent human factors work. The studies reported below followed directly from the questions produced by these data.

IV. CONTROL OF USER SPEECH

Having determined the speech characteristics of customers giving numbers to operators, we conducted a series of three studies designed to investigate methods of eliciting spoken numbers from users of automated network services that would be acceptable to ASR systems. These studies are summarized below.

4.1 Simulation study

The first study of the series involved a laboratory simulation of a connected-speech, speaker-independent ASR system in CCS. Each of the 30 subjects in the study was given a credit card with a 14-digit number printed with 3 3 4 4 segmentation. Subjects were then instructed to make a series of between 24 and 36 credit card calls that required voice entry of the credit card number to the simulated ASR system.

4.1.1 Prompts and feedback

After dialing a ten-digit number, most subjects heard a tone followed by a brief pause, then the following prompt:

> Please speak your credit card number in groups of four digits or less. Please wait for the numbers to be repeated back to you before speaking the next group. If the numbers are repeated incorrectly, say the word "error." Then when you hear the tone, repeat the last group of numbers spoken. Listen again for the response and then proceed with the next group of numbers. At the tone, please begin speaking your credit card number.

A similar prompt was used for another group of subjects that was not given feedback of the spoken digit groups. The instructions for this group included the phrase "say zero instead of oh." This phrase was not included when feedback was provided because we wanted to see the effect of feeding back "zero" when the subject said "oh." The subjects did not know that digit recognition was done by a human who keyed in the spoken segments of the number, triggering voice feedback of the spoken digits to those subjects selected to receive feedback.

We attempted to implicitly control speaking rate by means of the rate at which feedback was given; digits were fed back at either 2.50 or 1.25 digits per second. To investigate the effect of system recognition error rate on user input, two different digit recognition error rates of 1.8 percent and 5.4 percent were simulated. Each subject experienced at least two different combinations of feedback rate and error rate.

4.1.2 Summary of results

Subjects were initially told that once they thought they no longer needed to hear the prompt they could begin to speak after the first

tone. The typical subject would listen to the prompt on the first one or two calls, then begin to speak after the first tone on all subsequent calls. Despite the length of the prompt and the limited exposure to it, subjects had little difficulty following the instructions. All subjects in the feedback condition corrected errors smoothly and consistently. There were very few departures from the digit vocabulary (0.1 percent of all calls) and subjects nearly always used proper segmentation (99.8 percent of all calls).

While all subjects not given feedback followed the instruction to say "zero" instead of "oh," the feedback "zero" less successfully induced subjects who received feedback to say "zero". Tbe proportion of those subjects saying "zero" on the first call was 0.3; this increased to 0.7 by the twelfth call.

The mean speaking rate across all conditions was 2.58 digits per second. While this rate is lower than the rate of 3.43 digits per second observed in the field, the speaking rate did not vary significantly as a function of either feedback (none, slow, fast) or system error rate. Subjects in the feedback conditions did lower their speaking rate slightly (by 0.2 digit per second) when correcting a system recognition error, but then returned to the higher rate on the next call. Thus, under the conditions of this study, user speaking rate was not sensitive to feedback, either in terms of feedback rate or recognition error rate. However, it should be noted that the simulated-recognition error rates in this study were independent of a subject's speaking rate. In a real ASR system, the system error rate would increase with increasing user speaking rate, possibly making users more sensitive to feedback.

4.2 Telephone prompt study

The simulation study showed that some aspects of subjects' speech could be effectively controlled (at least in a laboratory setting) with simple prompting. However, the instructions in that study did not attempt to control speaking rate. Therefore, the next study concentrated on the speaking rate problem. Instead of bringing subjects into a laboratory stimulation to evaluate prompts, we chose to contact subjects by phone in their own work environments and ask for their cooperation in a very brief experiment that required them to simply say their own home phone numbers, including the area codes. Participants heard one of a set of recorded prompts and responded by speaking their numbers. All subjects were employees at Bell Laboratories in Holmdel, New Jersey. Spoken home telephone numbers provided a stronger test of our ability to control speaking rate, since a highly familiar number is likely to be spoken more rapidly than an unfamiliar one.

4.2.1 Candidate prompts

Table I gives the four components from which six prompts were composed. Three of the prompts began with Component 1 in Table I; one of these was completed by adding only Component 2, while the other two were completed by adding both Component 2 and either the first or second sentence of Component 3. The remaining tbree prompts were the same as the above three, except that Component 1 was omitted.

This set of six prompts allowed separate evaluation of tbe effect on subjects of knowing that they were speaking to a machine and the effect of specific instruction on how to speak. The first sentence of Component 3 in Table I (rate instruction) was designed simply to lower the speaking rate; the second (isolation instruction) was designed to elicit isolated speech. Absence of either of those two sentences from the prompt provided a control condition. A total of 60 subjects provided data on these prompts, 10 for each of the 6 prompts.

4.2.2 Summary of results

Table II gives the mean speaking rate for each prompt, corrected for intersegment pauses. A 3×2 analysis of variance showed that informing subjects that they were speaking to machines significantly lowered speaking rates ($F = 6.04$, $p < 0.03$). Also, specific instruction about how to speak significantly affected speaking rate ($F = 7.83$, $p < 0.01$). The interaction between the machine information and specific instruction conditions was not significant ($F = 0.51$).

One other instruction evaluated in an early phase of this study deserves mention. This instruction was worded as follows:

> At the tone, please speak your telephone number. Say "zero" instead of "oh." Speak at the following rate: [Recorded voice speaking, "One, two, three (pause) four, five, six"].

This instruction produced the lowest mean speaking rate (1.31 digits per second) of any prompt evaluated. Prior to playing the prompt, the experimenter told the subjects that they would speak to a machine, instead of including that information as part of the prompt.

As can be seen in Table II, even wben no machine information and no specific rate instruction is given, the speaking rate is still considerably lower than that observed in the field study. As in the simulation study, this probably occurred because the subjects knew they were

Table I—Components of telephone study prompts

1. This is a machine which recognizes human speech.
2. At the tone, please speak your telephone number. Say zero instead of oh.
3. Speak slowly and distinctly.
OR
Pause briefly after each digit.

Table II—Speaking rates in the telephone prompt study (digits/s)

	Machine Information	
Instruction	Absent	Present
Control	2.41	2.04
Rate	2.18	1.54
Isolation	1.56	1.36

participating in an experiment and were making a special effort to speak clearly. That special effort cannot be expected in an actual service environment. Nonetheless, the results clearly show that simple instructions can considerably lower the speaking rate.

Beyond the fact that the isolation instruction lowered user speaking rates, it is of interest to know if it also succeeded in eliciting isolated speech. Although the speech given in response to the instruction sounded adequately isolated to the listener, the data were not recorded in a form that allowed a more thorough treatment of the isolation question. This question is directly addressed in tbe next study.

4.3 Speech isolation study

Up to this point, the reported work has focused primarily on issues surrounding connected-speech recognition. For reasons related to a larger, ongoing ASR project, the focus of our human factors work now shifted to the problem of eliciting speech acceptable to an isolated speech recognition system. We wanted to see whether it was possible, through the use of prompts alone, to obtain isolated speech from subjects. While it is possible to force isolation through the use of a pacing cue or feedback after each spoken item, such techniques tend to produce slower input from the user than is required by the ASR system. Particularly in applications involving entry of long digit strings (such as credit card numbers), experienced users may find paced entry tedious. We adapted the procedure used in the previous study to the current needs. Besides attempting to develop in this study an initial prompt that would produce isolated speech, we also investigated the use of a reprompt to be used if the speech given in response to the initial prompt was not adequately isolated. Each of 90 subjects (60 Bell Laboratories employees and 30 from the surrounding community) heard one of a set of prompts over the telephone and responded with their home telephone numbers. Because of the concerns of the larger project, subjects were asked to give their numbers without the area codes.

4.3.1 The initial prompt

Based on the results of the previous study, we selected two variations

of each of two prompts for evaluation as the initial prompt. The candidates were:

1. At the tone, please say your number. Pause briefly between digits.
2. At the tone, please say your number. Pause briefly after each digit.
3. At the tone, please say your number. Pause between digits, like this. (Recorded voice speaking, "One, two, three")
4. At the tone, please say your number, as follows: (Recorded voice speaking "One, two, three")

The phrase "say zero instead of oh" was not used because reliable recognition of "oh" by an isolated speech system is not as difficult as it would be with a connected speech system, due to the absence of coarticulation effects. Also, in an attempt to keep the prompts as short as possible, we did not include a sentence telling customers that they would speak to a machine.

4.3.2 The reprompt

To evaluate the effect of a second attempt, a subject whose speaking time did not exceed 5.0 seconds on the first attempt received a second prompt, as follows:

We're sorry, would you please say your number again, but pause longer between digits.

4.3.3 Evaluation of isolation

Since the central question in this study was whether subjects were producing isolated speech, we needed a definitive test of isolation. This was provided by recording subjects' speech on analog tape and sending the tapes to the Bell Laboratories Acoustics Research Department at Murray Hill, New Jersey, where they were processed for end-point detection. Of primary interest were the number of isolated segments detected in each subject's speech. Since subjects spoke their own seven-digit home telephone numbers, seven segments should be detected on a number spoken with correctly isolated speech.

4.3.4 Summary of results

After gathering data from 40 subjects, 10 for each prompt, it was evident that Prompt 4 was unacceptable. Four of the ten subjects hearing this prompt responded with three digits or expressed confusion. Prompt 4 was therefore eliminated from further consideration. Data were then collected from an additional 30 subjects, 10 for each remaining prompt. After an initial evaluation of these data, 20 more subjects were added to better discriminate between Prompts 1 and 2.

For Prompt 1, 17 of the 30 subjects took longer than 5.0 seconds to say their telephone numbers on the first attempt and were therefore

not given the reprompt. The end-point detector found seven isolated segments in 14 of those 17 first attempts and six segments in each of the remaining two (one was missing from the tape). Of the 13 subjects given the reprompt, seven segments were detected on tbe first attempt for two subjects and on the second attempt for 12. The speech level was too low to segment on the second attempt of the remaining subject.

For Prompt 2, 22 of the 30 subjects were not given the reprompt. Of these, 18 achieved perfect isolation as determined by the end-point detector. Five segments were detected for two suhjects, six for anotber, and again one subject was missing from the tape. None of the eigbt suhjects given the reprompt spoke seven isolated segments on tbe first attempt. Six of the eight spoke perfectly isolated digits after the reprompt. One of the remaining two gave three segments and the other gave four.

For Prompt 3, 19 of the 20 subjects were not given tbe reprompt. Of these, 14 were determined to have spoken seven isolated digits. One spoke seven digits in six segments and another in five. The end-point detector found five digits in five segments for two suhjects, with the speech level on the remaining two digits being too low to segment. For a third suhject, there were six isolated digits, plus one with too low a level. The one suhject given the reprompt spoke six segments on the first attempt and seven on the second.

The results of this study are an encouraging indication that the combination of an initial prompt with a reprompt can be effective in eliciting isolated speech. The differences in effectiveness among the initial prompts are not large, hut are in favor of Prompt 3. However, since that prompt is considerahly longer than either of the others, use of one of the comparable shorter prompts is preferahle in any real application.

V. CONCLUSION

The work reported ahove represents the initial human factors steps toward eventual use of ASR in Bell System network applications. That effort has been concentrated on those human factors questions surrounding the user-system dialog. As indicated in Section II of this paper, there are many other human factors issues that must be faced hefore any network application of speaker-independent ASR will he possible. However, our work on those remaining issues is heyond the scope of this paper. Also heyond the present scope are those potential telecommunications applications of hoth speaker-dependent and speaker-independent ASR outside the network. This hroader range of applications raises several human factors questions in addition to those considered here.

VI. ACKNOWLEDGMENTS

Several people deserve thanks for their contributions to the work summarized here. H. Holinka assisted in all of the studies. L. Chapman, a visiting student from the University of Texas at El Paso, helped plan and collect data for the laboratory simulation. T. M. Gruenenfelder was responsible for analysis of the data from that study. M. L. Viets was heavily involved in all phases of the telephone prompt and speech isolation studies. C. J. Karhan served as primary coordinator for the project of which the speech isolation study was a part. I would also like to thank L. R. Rabiner and J. G. Wilpon of the Acoustics Research Department at Bell Laboratories for evaluating the isolation of the recorded speech. Finally, special thanks are due E. A. Youngs, who initiated the human factors effort reported here and has been a constant source of ideas and encouragement.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE, *64* (April 1976), pp. 487–501.
2. J. M. Baker, "How to Achieve Recognition: A Tutorial/Status Report on Automatic Speech Recognition," Speech Technology, *1*, No. 1 (Fall 1981), pp. 30–43.
3. C. S. Meyers and L. R. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition," B.S.T.J., *60*, No. 7, Part 2 (September 1981), pp. 1389–1409.
4. A. E. Rosenberg, L. R. Rabiner, and J. G. Wilpon, "Recognition of Spoken Spelled Names for Directory Assistance Using Speaker-Independent Templates," B.S.T.J., *59*, No. 4 (April 1980), pp. 571–92.
5. S. E. Levinson, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 2 (April 1979), pp. 134–41.
6. L. R. Rabiner, private communication.

AUTHOR

John E. Holmgren, B. S. (Psychology), 1965, University of Wisconsin; Ph.D. (Mathematical Psychology), 1970, Stanford University; Bell Laboratories, 1979–1982; American Bell, 1982—. In the Human Factors Department at Bell Laboratories, Mr. Holmgren worked on potential applications of automatic speech recognition. His other major responsibility was in the area of audiographics teleconferencing. In July 1982, Mr. Holmgren transferred to American Bell, where he is a Group Supervisor in AIS Net 1000 Research and Development. Member, Psychonomic Society.